DOCUMENT RESUME

ED 331 888 TM 016 483

AUTHOR Wise, Steven L.; And Others
TITLE A Comparison of Self-Adapted and Computer-Adaptive Tests.
PUB DATE Apr 91
NOTE 20p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 3-7, 1991).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Ability Identification; *Adaptive Testing; *College Students; Comparative Testing; *Computer Assisted Testing; Difficulty Level; Higher Education; Item Banks; *Item Response Theory; Statistics; *Test Anxiety
IDENTIFIERS *Self Adapted Testing

ABSTRACT

According to item response theory (IRT), examinee ability estimation is independent of the particular set of test items administered from a calibrated pool. Although the most popular application of this feature of IRT is computerized adaptive (CA) testing, a recently proposed alternative is self-adapted (SA) testing, in which examinees choose the difficulty level of each of their test items. Examinee performance was compared under CA and SA testing conditions for college students from an introductory statistics course. Three test forms were developed, testing mathematical knowledge necessary for the course. The final pool contained 93 items which were administered to 204 subjects. The SA test yielded significantly higher ability scores, and examinees taking the SA test reported significantly lower posttest state anxiety. Implications of the differences between the two test types for measurement practice are discussed. Three tables present study data. (Author/SLD)

A Comparison of Self-Adapted and Computer-Adaptive Tests

Steven L. Wise, Barbara S. Plake,

Phillip L. Johnson, and Linda L. Roos

University of Nebraska-Lincoln

Paper presented at the 1991 annual meeting of the

American Educational Research Association, Chicago, IL

Abstract

According to item response theory (IRT), examinee ability estimation is independent of the particular set of test items administered from a calibrated pool. Although the most popular application of this feature of IRT is computerized adaptive (CA) testing, a recently proposed alternative is self-adapted (SA) testing, in which examinees choose the difficulty level of each of their test items. This study compared examinee performance under SA and CA tests, finding that the SA test yielded significantly higher ability scores and examinees taking the SA test reported significantly lower post-test state anxiety. Implications of the differences between the two test types for measurement practice are discussed.

A Comparison of Self Adapted and Computerized Adaptive Tests

The development of item response theory (IRT) has made it possible for the test performance of examinees to be compared on the same scale of measurement even if they are administered different sets of test items. According to IRT, examinee ability estimation is independent of the particular set of test items administered from a calibrated pool (Hambleton & Swaminathan, 1985). The most popular application of this feature of IRT is computerized adaptive (CA) testing, in which a computer algorithm is used to match the difficulty levels of the items administered to the ability level of each examinee. At each step in the CA testing process, the next item to be administered is a function of the examinee's responses to items previously administered. Examinee characteristics such as test anxiety and motivation are not taken into account during CA testing.

Rocklin and Thompson (1985) found that, when administered a relatively difficult test, examinees lower in test anxiety tended to perform better than their more anxious peers. When a relatively easy test was administered, however, examinees reporting moderate levels of test anxiety performed better than examinees reporting either lower or higher levels of anxiety. On the basis of these findings, Rocklin and O'Donnell (1987) explored the effects of allowing examinees to choose the difficulty levels of their items on a computer-based test. Their procedure, termed self-adapted testing, allows an examinee to choose the difficulty of each item to be administered from several levels of difficulty. Rocklin and O'Donnell (1987) differentiated self-adapted (SA) testing from computerized adaptive (CA) testing in the following way: "instead of being

tailored to the examinee's estimated ability level, a self-adapted test is tailored to the examinee's self-perceived ability as well as to his or her current motivational and affective characteristics" (p. 315). In their study, Rocklin and O'Donnell calibrated a pool of 40 items using a one-parameter (Rasch) IRT model. Examinees were randomly assigned to be administered either (a) a relatively easy test, consisting of the 20 least difficult items in the pool, (b) a relatively difficult test, consisting of the 20 most difficult items, or (c) a 20-item SA test, in which examinees chose the difficulty level of each item before it was administered. The results showed that the SA test yielded a significantly higher mean ability score than the two traditional tests. Moreover, the three tests did not differ significantly in terms of standard error of ability estimation.

Rocklin and O'Donnell (1987) interpreted the higher scores on the SA test as an indication that examinees were able to make effective and strategic choices among the items, suggesting that "an examinee has access to a variety of information (including current affective and motivational states) relevant to optimal item selection beyond that which would be available to a traditional computerized adaptive testing algorithm" (p. 318). Note, however, that their study did not explicitly compare SA and CA tests.

The purpose of the present study was to compare the relative effects of SA and CA tests on examinee performance. Specifically, comparisons were made in terms of (a) estimated ability, (b) post-test anxiety level, (c) total testing time, and (d) variance error of ability estimation.

Method

Subjects

The subjects were 204 students from five sections of an introductory statistical methods course at a large midwestern university. The group of

subjects consisted of 156 undergraduates and 48 graduate students. There were 76 males and 128 females in the sample. Participation in the study was a required part of the statistics course; the test results were used to identify students in need of remediation in basic algebra skills. Subjects were randomly assigned to the two testing conditions used in the study.

Prior to data collection, a power analysis was used to choose a sample size that would yield an experimental design with adequate sensitivity to detect meaningful-sized treatment effects. Considering that the IRT-based ability scale would have a standard deviation of approximately one, it was judged that the smallest meaningful mean difference between the two test types would be .25 points. Using Cohen's (1988) tables, this corresponded to a standardized effect size of .25. Because, however, a blocking variable was used, it was anticipated that the standardized effect size would increase to roughly .35 due to the consequent decrease in error variance. Using a .05 significance level and expecting 100 subjects per test type, this effect size corresponded to a power value of .68, which was deemed adequate for this study.

Item Pool Development

A pool of items was developed to be used in identifying those introductory statistics students whose basic mathematics skills were in need of remediation. Initially, a pool of 120 items was developed. Each item used a four-choice multiple-choice format. The content of the items addressed the types of mathematics skills that are needed during a course in introductory statistical methods. The majority of the items concerned algebra skills such as basic operations, solving equations, order of operations, inequalities, and linear equations. A small number of additional items dealt with probability and logical reasoning.

After the initial item pool was developed, two of the authors and two graduate students independently rated each item on (a) its acceptability as a measure of a mathematics skill needed for introductory statistics, and (b) its difficulty level. On the basis of these ratings, unacceptable items were deleted/modified and three 35-item test forms were constructed. These forms were similar in terms of both difficulty (based on the item ratings) and item content. Each item appeared on only one test form resulting in 105 unique items across the forms.

The three test forms were randomly administered to all students taking an introductory statistics course at a midwestern university between January, 1988 and July, 1989. The tests were delivered on Apple IIe microcomputers using programs written in Applesoft BASIC. Approximately 250 examinees were administered each test form. To evaluate the dimensionality of the item pool, the data from each test form were analyzed using a principal-axis factor analysis. In each factor analysis, a single factor was extracted and items with factor loadings less than .10 were deleted from the pool. This procedure resulted in the deletion of 12 items, leaving a final pool of 93 items. After items with low loadings were deleted, the first eigenvalue accounted for between 20 and 25 percent of the total variance in each test form. According to the criteria proposed by Reckase (1979), a test can be considered sufficiently unidimensional for unifactor IRT if a factor analysis yields a first eigenvalue accounting at least 20 percent of the variance. Note that because each examinee was administered only one test form, it was not possible to directly assess the dimensionality of the entire item pool. Considering the procedures used to construct the three similar test forms, however, and given the evidence

of dominant first factors in each test form, it was inferred by the authors that the item pool was sufficiently unidimensional to apply IRT.

The 93 items in the final pool were calibrated using the LOGIST computer program (Wingersky, Barton, & Lord, 1982). A modified one-parameter logistic model was used in which the lower asymptote of each item characteristic curve was set at .20. Barnes and Wise (in press) found that, for smaller sample sizes (no more than 200 examinees), the modified one-parameter model was more effective in estimating item parameters than either the one-parameter or three-parameter models.

Instruments

The SA and CA tests were both administered on IBM microcomputers using the MicroCAT testing software (Assessment Systems Corporation, 1988). An IRT ability score for each examinee was calculated using maximum-likelihood estimation. In addition, the MicroCAT program computed the variance error of ability and the total testing time. Each test used a fixed length of 20 items.

In developing the SA test, the 93 item difficulty (b) parameters were ranked and divided into six difficulty levels, with each level containing 15 or 16 items. Next, the items within each level were randomly ordered and a MicroCAT specification program was developed that "fixed" the administration of items to that randomly-determined order. That is, all examinees choosing a particular level were administered the same first item, the same second item, and so on.

At the beginning of the SA test, examinees were presented the following instructions:

> This 20-item test is intended to measure your level of proficiency in the types of mathematics skills that are needed for a course in introductory statistics. This test is different from most tests that you have taken. Before each test item is presented, you will choose how difficult you want the item to be. You will choose among six different levels of difficulty, ranging from level 1 (the easiest items) to level 6 (the hardest items).
>
> The higher the difficulty level of an item that you choose, the more credit you will receive if you pass that item. When calculating your score on this test, we will take into account the difficulty levels of the items you have chosen, and credit your answers accordingly.
>
> We recommend that you choose the hardest items that you think that you can answer correctly. You are, however, free to choose whatever item difficulty levels that you prefer. The items are weighted in such a way that it should not matter which items you have chosen -- your final score should be about the same.

After an examinee answered a given item, feedback was given to the examinee in the form of a message stating which lettered option was correct, and the examinee was then asked to choose the difficulty level of the next item. Because there were fewer than 20 items at each level, examinees sometimes chose levels exhausted of items. In these instances, examinees were informed that no more items were available at that level and they were asked to choose again.

Examinees taking the CA test received the following instructions:

> This 20-item test is intended to measure your level of proficiency in the types of mathematics skills that are needed for a course in introductory statistics. This test is different from most tests that you have taken. The items that you receive are chosen by the computer based on your performance. That is, every time you pass an item, you'll be given a more difficult item; every time you fail an item, you'll be given an easier item. The computer will take into account the difficulty levels of your items when calculating your score on the test.

Feedback was also provided after each item to examinees taking the CA test.

In both tests, when 20 items had been administered to an examinee, his or her ability estimate was calculated. Depending on the ability score, the examinee was informed whether he or she was required to attend a one-hour mathematics remediation session to be held the following week. An ability estimate cutoff value of .20 was chosen to be used in deciding whether or not an examinee required remediation. Based on the test characteristic curve for the entire item pool, an ability score of .20 corresponded to a domain score of .77.

An additional instrument was used in this study. The State Anxiety Scale of the State-Trait Anxiety Inventory (Spielberger, Gorsuch, & Lushene, 1970) was used to measure examinee situation-specific anxiety before and after administration of the mathematics test. The manual for the State Anxiety Scale cites strong evidence of the instrument's reliability and validity.

Procedure

The testing was conducted during the first week of class. During the first class session, students (a) were informed that the test scores would be used to

identify students in need of mathematics remediation, (b) completed a demographic sheet in which they provided information regarding the number of algebra courses they had taken and the number of years since their last algebra course, and (c) signed up for a time to be administered the mathematics test.

The mathematics test was administered during the first week of class in large, quiet room containing 10 identical IBM PS/2 Model 55 microcomputers. Five of the microcomputers were set up to administer the SA test, with the remainder administering the CA test. Halfway through the week of testing, the microcomputers administering the SA test were switched to administer the CA test, and vice versa. Students were assigned to test type through their self-selection of a microcomputer upon arrival for testing.

Students were tested in groups ranging in size from one to nine. When each student arrived, the test administrator directed the student to choose a microcomputer and to complete a paper-and-pencil version of the State Anxiety Scale. Next, the student completed the computer-based mathematics test. Pencils and scratch paper were provided and the use of calculators was not allowed. No time limit was imposed during testing. Upon completion of the test the students were informed, via the computer, whether or not they were required to attend a remediation session, dependent on their ability estimates. Finally, the State Anxiety Scale was again administered.

Data Analysis

There were four dependent variables of interest in this study: (a) estimated ability, (b) post-test anxiety, (c) total testing time, and (d) variance error of estimated ability. The primary independent variable was test type (SA, CA). In the analyses involving estimated ability and post-test state anxiety, however, post hoc blocking variables were used to increase the sensitivity of

the design through reduction of error variance. For estimated ability, a blocking variable of the number of years since last algebra course was employed. Wise, Plake, Eastman, Boettcher, and Lukin (1986) found, in a similar testing context, that the number of years since last algebra course was substantially related to test performance. Three blocks were formed to yield nearly equal sample sizes. The three blocks corresponded to (a) less than three, (b) three to five, or (c) more than five years since last algebra course. For post-test state anxiety, pre-test state anxiety was employed as a blocking variable. Three blocks, formed to yield nearly equal sample size within blocks, were defined by the following score ranges: (a) less than 33 (Low), (b) 33-41 (Medium), and (c) greater than 41 (High). The blocking variable levels (Low, Medium, High) were derived from the distribution of State Anxiety Scale scores found in this study. Examinees in the high anxiety group, for example, scored relatively high in this study, but not necessarily high according to the State Anxiety Scale norms.

The data for estimated ability and post-test state anxiety were each analyzed using a two-factor analysis of variance (ANOVA). Because the distributions of scores for total testing time and variance error of ability showed a marked degree of skewness, a nonparametric test was appropriate for comparing the two test types. The large-sample z approximation of the Mann-Whitney U test was used (Hays, 1981). A .05 level of significance was used in all analyses.

Results

Means and standard deviations for estimated ability are shown in Table 1. The ANOVA for this dependent variable found a significant effect for both the blocking variable ($F(2,198)=16.11$, $MS_e=1.05$, $p<.001$) and test type ($F(1,98)=5.19$, $MS_e=1.05$, $p=.024$). The interaction between the blocking

variable and test type was nonsignificant. The mean ability score for examinees who took the SA test was significantly higher than for those taking the CA test.

Insert Tables 1 and 2 about here

Table 2 contains the means and standard deviations for post-test anxiety. The results of the ANOVA showed that the effect for the blocking variable was significant ($F(2,198)=63.23$, $MS_e=84.45$, $p<.001$) as was the effect for test type ($F(1,198)=4.16$, $MS_e=84.45$, $p=.043$). The interaction effect was found to be nonsignificant. Examinees taking the SA test reported significantly lower mean post-test state anxiety than those taking the CA test.

As shown in Table 3, the distributions of testing time and variance error of ability were substantially skewed. In terms of testing time, there was a median difference of about three and one half minutes between the two tests. The results of the Mann-Whitney U test indicated that examinees taking the SA test took significantly more time to complete the test ($z=2.18$, $p=.029$). Moreover, the variance error of ability was found to be significantly larger for the SA test ($z=3.60$, $p<.001$).

Insert Table 3 about here

Discussion

The finding that examinees taking the SA test scored significantly higher than those taking the CA test, is consistent with the results found by Rocklin and O'Donnell (1987). In addition, examinees reported significantly less anxiety after testing than examinees taking the CA test. The CA test yielded

significantly more precise ability estimates; however, this result was not surprising, because the CA testing algorithm specifically chose the items that most lowered the variance error of ability estimation. The finding that total testing time was significantly longer for the SA test is also not surprising, since a computer is likely to choose the next item to be administered much more rapidly than will an examinee.

The results of this study are intriguing. The logic underlying CA testing requires the assumption that ability estimation is independent of the items that are administered. From this assumption, it follows that the two test types should not have differed in terms of examinee performance. The findings of Rocklin and Thompson (1985), Rocklin and O'Donnell (1987), and the present study are, however, contrary to the assumption.

There are a number of questions regarding SA testing that are in need of further investigation. One major question concerns the mechanism by which SA testing enhances performance over traditional testing methods. The results of this study suggest that, for examinees taking the SA test, higher test performance was related to lower anxiety levels. The nature of this relationship, however, is not clear. Does lowered anxiety serve to enhance test performance or does higher test performance (and accompanying positive item feedback) tend to lower anxiety? Secondly, how important is item feedback to the success of SA testing? Rocklin (1989) suggested that item feedback is important on a SA test because it allows examinees to make more informed choices of subsequent item difficulty levels. It may be the case, however, that SA testing will improve examinee test performance even when feedback is absent. Finally, to what extent is the higher test performance resulting from use of a SA test due to a "novelty" effect that may subside with additional use of this testing format?

That is, do examinees perform higher on a SA test only because they are more motivated by its non-traditional testing format? Answers to these questions will help researchers to better understand how SA testing influences test performance. If SA testing lowers examinee test anxiety and consequently leads to higher test performance, then this novel testing format represents a new application of IRT that should be quite useful to measurement practitioners.

References

Assessment Systems Corporation (1988). User's manual for the MicroCAT Testing System, Version 3. St. Paul, MN: Author.

Barnes, L. L. B., & Wise S. L. (in press). The utility of a modified one-parameter IRT model with small samples. Applied Measurement in Education.

Hays, W. L. (1981). Statistics (3rd. Ed.). NewYork: Holt, Rinehart and Winston.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston: Kluwer-Nijhoff.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. Journal of Educational Statistics, 4(3), 207-230.

Rocklin, T. (1989, March). Individual differences in item selection in computerized self adapted testing. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Rocklin, T., & O'Donnell, A. M. (1987). Self-adapted testing: A performance-improving variant of computerized adaptive testing. Journal of Educational Psychology, 79(3), 315-319.

Wingersky, M. S., Barton, M. A., & Lord F. M. (1982). LOGIST user's guide. Princeton, NJ: Educational Testing Service.

Wise, S. L., Plake, B. S., Eastman. L. A., Boettcher, L. L., & Lukin, M. E. (1986). The effects of item feedback and examinee control on test performance and anxiety in a computer-administered test. Computers in Human Behavior, 2(1), 21-29.

Author Notes

The authors wish to thank Leslie Lukin, Laura Barnes, and Bunny Pozehl for their assistance in developing the item pool and Thomas Rocklin for providing the self-adapted testing MicroCAT code that was modified for use in this study.

Correspondence regarding this article should be sent to Steven L. Wise, Department of Educational Psychology, 122 Bancroft Hall, University of Nebraska, Lincoln, NE 68588-0345.

Table 1

Descriptive Statistics for Estimated Ability, By Test Type and Years Since Last Algebra Course

	Test Type					
	SA			CA		
Years Since Last Algebra Course	Mean	SD	n	Mean	SD	n
Less Than Three	0.84	0.72	44	0.43	1.00	45
Three to Five	0.32	1.30	27	-0.10	1.06	33
More than Five	-0.27	1.17	31	-0.44	0.91	24
All Examinees	0.37	1.14	102	0.06	1.05	102

Table 2

Descriptive Statistics for Post-Test State Anxiety, By Test Type and Pre-Test State Anxiety Level

	Test Type					
	SA			CA		
Pre-Test State Anxiety Level	Mean	SD	n	Mean	SD	n
Low	28.30	6.20	30	30.54	9.92	35
Medium	35.60	8.00	35	39.22	10.87	32
High	46.05	10.35	37	48.09	8.69	35
All Examinees	37.25	11.13	102	39.28	12.17	102

Table 3

Descriptive Statistics for Total Testing Time and Variance Error of Ability

Dependent Variable	Test Type	Minimum	Median	Maximum
Testing Time (Minutes)				
	SA	10.23	22.15	67.82
	CA	7.72	18.70	52.45
Variance Error of Ability				
	SA	0.10	0.14	22.08
	CA	0.10	0.12	3.48